# Emotional Sensitivity in Human-Computer Interaction

## Emotionale Sensitivität in der Mensch-Maschine Interaktion

Jonghwa Kim, Johannes Wagner, Thurid Vogt, Elisabeth André, Frank Jung, Matthias Rehm,
Universität Augsburg

**Summary** Human conversational partners usually try to interpret the speaker's or listener's affective cues and respond to them accordingly. Recently, the modelling and simulation of such behaviours has been recognized as an essential factor for more natural man-machine communication. The implicit emotion channels of human communication such as speech, facial expression, gesture, and physiological responses are used in general to extract emotion-relevant features for the computational perception of emotion. So far, research on emotion recognition has mostly dealt with offline analysis of recorded emotion corpora, and online processing (in realtime or near realtime) has hardly been addressed. Online processing is, however, a necessary prerequisite for the realization of human-computer interfaces that analyze and respond to the user's emotions while he or she is interacting with an application. In this paper, we first describe how we recognize emotions from various modalities including speech, gestures and biosignals. We then present Smart Sensor Integration (SSI), a framework which we developed to meet the specific requirements of online emotion recognition. ►►► **Zusammenfassung** Menschliche Gesprächspartner versuchen in der Regel, emotionale Regungen ihres Gegenübers zu erkennen und entsprechend darauf zu reagieren. Mit dem Ziel, die Mensch-Maschine Interaktion einfacher und natürlicher zu gestalten, beschäftigen sich zunehmend Forschergruppen damit, emotionales Verhalten maschinell zu simulieren. Zur maschinellen Erkennung der Emotion werden die impliziten Kommunikationskanäle menschlicher Emotion wie z. B. Sprache, Gesichtsausdruck, Gesten und physiologische Reaktionen am häufigsten verwendet, um affektive Merkmale zu extrahieren. Allerdings konzentrierte sich die Forschung bisher auf die Offline-Analyse emotionaler Korpora, und eine Online-Verarbeitung (in Echtzeit oder Nah-Echtzeit) wurde selten angestrebt. Eine Integration von Methoden zur Erkennung von Emotionen in praktische Anwendungen setzt jedoch voraus, dass ein System in der Lage ist, auf Emotionen eines menschlichen Nutzers zu reagieren, während dieser mit dem System interagiert. In diesem Papier skizzieren wir zunächst, wie Emotionen aus Sprache, Gestik und Biosignalen erkannt werden können. Anschließend stellen wir Smart Sensor Integration (SSI) vor, ein Werkzeug, das auf die besonderen Bedürfnisse der Online-Emotionserkennung zugeschnitten ist.

**Keywords** H.5 [Information Systems: Information Interfaces and Presentation]; I.2 [Computing Methodologies: Artificial Intelligence]; Emotion recognition, affective computing, multimodal fusion, multimodal user interfaces, speech recognition, physiological signal ►►► **Schlagwörter** Emotionserkennung, multimodale Fusion, multimodale Benutzerschnittstellen, Spracherkennung, physiologische Signale

## 1 Introduction

During the last decade, burgeoning interest in achieving emotional sensitivity in machines has been prompted in human-computer interaction (HCI). Methods have been developed to detect a user's emotions from various modalities including facial expressions [12], gestures [2], speech [8], and physiological measurements [5]. Also, multimodal approaches to improve recognition accuracy are reported, mostly by exploiting audiovisual combinations [1]. There is empirical evidence that many problems in man-machine communication could be avoided if the machine was more sensitive towards a user's feelings.

Typical examples include dialogues with non-human operators in call centers.

One of the most decisive hurdles in bringing emotion recognition systems into practical applications is the fact that most systems so far have been developed for offline processing. This is due to varied difficulties realtime capability implies, e. g., requirements for automatic segmentation and low-cost algorithms. Apart from that it is no longer possible to handle processing in temporally independent steps, which also hampers the use of specialized tools for each task. Instead, simultaneous execution is required, i. e., sensor data must be permanently captured and processed, while at the same time classification has to be invoked on detected segments. This becomes even more complex for multimodal input which requires support for fusing data from different sensors.

In the following, we describe our attempts to move from offline to online emotion recognition. First, we describe a general pipeline for multimodal emotion recognition. We then move to modality-specific recognition processes focusing on speech, gestures, and physiological data. After that, we describe a framework for multimodal processing in realtime, called Smart Sensor Integration (SSI), which we developed to jump-start the development of multimodal emotion recognition systems for online applications.

## 2 Methodology

Aside from considering certain system-specific components, architectures of offline and online emotion recognition systems reduce in general to three main tasks;

(a) *Data segmentation* is the task of detecting meaningful actions in the signal. In online systems the units of choice are usually well-defined segments expressed under certain emotions. In an online system start and end have to be detected automatically and emotional changes may happen throughout a term.

(b) *Feature extraction* relates to the task of finding meaningful features from the signal. This can include several steps, covering from pre-processing of the raw signal over the calculation of low-level features to the extraction of high-level statistics. For online systems, the choice of features is restricted to those that can be calculated possibly in realtime or near realtime at least. Therefore the effective use of feature selection and realtime signal processing techniques plays an important role.

(c) *Classification* is mapping observed feature vectors to discrete emotional states, such as joy, anger, sadness or surprise, or to continuous values of a dimensional emotional model, such as the PAD model (using pleasure, arousal, dominance axes) by Mehrabian [6]. Labeling emotional data with discrete states, most research is based on supervised pattern recognition approaches. The recognition performance of such approaches will be carried away by the quality (and the quantity) of the training data set. For online systems, the classifier can be trained in advance with available data sets under offline condition (hidden training), or gradually (incremental training) by using data sets that are cumulatively available during the online use.

For a multimodal approach, a fourth task has to be accomplished: the fusion of different sources. Generally, fusion can take place at three levels: data-level by merging raw sensor data, feature-level by merging multiple feature vectors to a single vector, and decision-level by combining decisions obtained from each modality.

## 3 Modalities

### 3.1 Speech

Information on emotion is encoded in all aspects of language, in what we say and in how we say it, and the 'how' is even more important than the 'what'. Focusing on the phonetic and acoustic properties of affective spoken language, we analyze/recognize emotions offline as well as online by using our EmoVoice system [9].

Finding meaningful speech units suitable for emotion classification is the first task of the system. We employ automatic *voice activity detection* to segment the incoming signal into chunks of voice activity, which roughly coincide with linguistic phrase boundaries. This method makes us independent from automatic speech recognition (ASR) that is needed for online segmentation into phrases, but may be faulty and time-consuming.

The acoustic features we used are mainly based on short-term acoustic observations, including pitch, signal energy, Mel-frequency cepstral coefficients, spectral and voicing information, and the harmonics-to-noise ratio. All values of an observation stream are transformed to different views (only local maxima, for example), and for each of the resulting streams 9 statistical values are calculated. Furthermore, a few segment-level durational and voice quality features are added. For the online system, we use only features that are fully automatically extractable in realtime which is opposed to many offline approaches to speech emotion recognition that rely to some extent on manually annotated information. Overall, we thus obtain a set with 1451 features which is reduced to about 20–150 features by feature selection.

For testing the performance of the system, we evaluated an acted database that is commonly used in offline research (7 emotion classes, 10 professional actors) and a speech database we recorded for online classification (4 emotion classes, 10 German students) and achieved an average recognition accuracy of 80% and 41% respectively.

### 3.2 Gestures

To capture a user's gestural behavior, we rely on acceleration sensors by the Wiimote (Wii™). Gallaher [3] categorizes gestural style by a number of expressivity parameters, e. g., how fast a gesture is done, how much space one uses to perform a gesture etc. Taking these param-

eters as indicators of a user's affective state, expressivity recognition with the Wiimote is defined as a two-class classification problem (Low, High) for each parameter. To test the feasibility of this approach, the parameters power, speed, and spatial extent were chosen. Three classifiers were used for this task, one for each parameter that was trained on the two-class problem of distinguishing between low and high values for the expressivity parameters. In a first step, features were calculated on the raw signal. For the acceleration data, we calculated the length of the signal, the minimum and maximum for each axis, the median and mean for each axis, and the gradient for each axis. The training set consists of 1260 samples by 7 subjects, i. e., 420 samples for each expressivity parameter (210 samples per class). A ten-fold cross-validation of a Nearest Neighbor classifier yielded recognition results of at least 94% for power, speed and spatial extent [7]. In collaboration with partners from the EU project CALLAS, we are currently preparing an experiment to analyze correlations between emotions and gestural expressivity using computer vision and data gloves in addition to the Wiimote.

### 3.3 Biosignals

As a further carrier of emotional information, we consider physiological data, including electrocardiography (ECG), electromyography (EMG), Galvanic skin response (GSR), and respiration (RSP). Unlike audiovisual signs, physiological reactions of a person are not directly observable by other humans and hence are primarily not used to communicate emotions. Instead, they serve as a biological process to control our behaviour in certain situations, e. g., to prepare our body to attack or escape from an enemy. Controlled by the autonomous nervous system the physiological reactions of the body are less subdued by the human will and to social masking, moreover they are permanently present and can be captured continuously.

To learn more about the mapping between the observed patterns and certain emotional states, we conducted several experiments during which we captured the biosignals from users while they were put into different affective states. In one, where music was used to elicit the desired emotions, we were able to distinguish four emotional states (joy, anger, sadness, and pleasure) with an accuracy of over 90% [10]. Recently we could achieve similar results by using a generic set of recursively calculated realtime features based on the same database [4].

In offline condition, we have a variety of choices for applying signal analysis techniques to obtain relevant features. In [5], we proposed a wide range of physiological features from various analysis domains, including time/frequency, entropy, geometry, subband spectra, etc., to search for the best emotion-relevant features and to correlate them with emotional states. Based on the best features we developed an online in-vehicle emotive monitoring system within the EU project METABO.

## 4 Smart Sensor Integration

As mentioned earlier, research so far has mostly dealt with offline evaluation of emotions, and online processing has hardly been addressed. In response to this we have developed a publically available framework, called Smart Sensor Integration (SSI)[1], which considerably jump-starts the development of online multimodal emotion recognition systems.

In the first place SSI offers an abstract interface to plug sensor devices in a pipeline like manner with dedicated processing modules, which permanently run on the captured data and store the transformed outcome to disk or share it with external applications, e. g., via sockets. However, a large number of signal processing algorithms have already been incorporated and can be directly accessed by developers. Through a multi-threaded design, several processing pipelines can run in parallel fed by different modalities, which are processed in a synchronized manner and may be combined at different fusion stages. In addition, external tools, such as OpenCV[2], a library for image processing, ARToolKitPlus[3], a library for marker tracking, and SHORE[4], a library for facial emotion detection, are also integrated. To support the building of user models from training data, a GUI has been developed, which runs on top of SSI and supports the complete machine learning setup starting from data acquisition and annotation, over feature extraction and training, to offline and online evaluation of the learned models. In all, the GUI paves the way for collecting multimodal corpora and analyzing each channel separately as well as in combination with other channels.

SSI has been successfully applied in a number of HCI projects at our lab, such as an attentive virtual butler which responds to the user's emotional state. Recently SSI is also being used in several EU-funded projects, e. g., in the METABO[5] for real-time physiological data analysis of diabetes patients in an automotive environment and in the CALLAS[6] and the Network of Excellence IRIS[7] for emotional multimodal interaction in artistic installations and story telling environments.

## 5 Conclusion

To achieve emotional sensitivity in machines is one of the hardest tasks in man-machine communication. In this paper, we presented various multimodal approaches to automatic emotion recognition and discussed challenges that arise when moving from offline emotion recognition to online emotion recognition where realtime constraints have to be met. As a practical guideline

[1] http://mm-werkstatt.informatik.uni-augsburg.de/ssi.html
[2] http://sourceforge.net/projects/opencvlibrary
[3] http://studierstube.icg.tu-graz.ac.at/handheld_ar/artoolkitplus.php
[4] http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/biom/dd.jsp
[5] www.metabo-eu.org
[6] www.callas-newmedia.eu
[7] http://iris.scm.tees.ac.uk

to online implementation, we introduced Smart Sensor Integration, a middleware which supports the development of multimodal emotional interfaces controlled by one or more sensors in realtime.

Challenging issue in the near future would be to assemble potential methods for computing human-like multimodal decision making and cognitive process in various contexts. For this, it requires not only a methodological, technical innovation but also conceptual changes with workable thoughts focusing on the context of applications.

## Acknowledgements

The work described in this paper is funded by the EU under research grants CALLAS[6] (IST-34800), IRIS[7] (Reference: 231824) and Metabo[5] (Reference: 216270).

## References

[1] Bailenson, J., Pontikakis, E., Mauss, I., Gross, J., Jabon, M., Hutcherson, C., Nass, C., John, C.: *Real-time classification of evoked emotions using facial feature tracking and physiological responses.* In: Int'l Journal of Human-Computer Studies, 66(5):303–317, 2008.

[2] Caridakis, G., Raouzaiou, A., Karpouzis, K., Kollias, S.: *Synthesizing gesture expressivity based on real sequences.* In: Proc. of LREC Workshop on multimodal corpora: from multimodal behaviour theories to usable models, 2006.

[3] Gallaher, P. E.: *Individual differences in nonverbal behavior: Dimensions of style.* In: Journal of Personality and Social Psychology, 63(1):133–145, 1992.

[4] Hönig, F., Wagner, J., Batliner, A., Nöth, E.: *Classification of user states with physiological signals: On-line generic features vs. specialized feature sets.* In: Proc. of the 17th European Signal Processing Conf. (EUSIPCO-2009), 2009.

[5] Kim, J., André, E.: *Emotion recognition based on physiological changes in music listening.* In: IEEE Trans. Pattern Anal. and Machine Intell., 30(12):2067–2083, 2008.

[6] Mehrabian, A.: *Framework for a comprehensive description and measurement of emotional states.* In: Genetic, social, and general psychology monographs, 121(3):339–361, Aug 1995.

[7] Rehm, M., Bee, N., André, E.: *Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions.* In: BCS-HCI '08: Proc. of the 22nd British HCI Group Annual Conf. on HCI 2008, pp. 13–22, Swinton, 2008.

[8] Vogt, T., André, E., Wagner, J.: *Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation.* In: Affect and Emotion in Human-Computer Interaction, vol. 4868, *LNCS*, Springer-Heidelberg, 2008.

[9] Vogt, T., André, E., Bee, N.: *A framework for online recognition of emotions from voice.* In: Proc. of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Kloster Irsee, Germany, 2008.

[10] Wagner, J., Kim, J., André, E.: *From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification.* In: Proc. IEEE ICME 2005, pp. 940–943, Amsterdam, 2005.

[11] Wagner, J., André, E., Jung, F.: *Smart sensor integration: A framework for multimodal emotion recognition in real-time.* In: Proc. of Affective Computing and Intelligent Interaction (ACII 2009), Sep 10–12, 2009.

[12] Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S.: *A survey of affect recognition methods: Audio, visual, and spontaneous expressions.* IEEE Trans. Pattern Anal. Mach. Intell., 31(1):39–58, 2009.

Received: September 19, 2009

**Dr. Jonghwa Kim,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: kim@informatik.uni-augsburg.de

**M.Sc. Johannes Wagner,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: johannes.wagner@informatik.uni-augsburg.de

**Dipl.-Inform. Thurid Vogt,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: thurid.vogt@informatik.uni-augsburg.de

**Prof. Dr. Elisabeth André,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: andre@informatik.uni-augsburg.de

**M.Sc. Frank Jung,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: frank.jung@informatik.uni-augsburg.de

**Priv.-Doz. Dr. Matthias Rehm,** Institut für Informatik, Universität Augsburg, Eichleitnerstr. 30, D-86159 Augsburg, Germany,
e-mail: matthias.rehm@informatik.uni-augsburg.de

The authors are working in the field of human-computer interaction and affective computing at the Chair of Multimedia Concepts and their Applications (www.interactive-multimedia.de), Augsburg University. They welcome comments and questions on their research results and make the presented Smart Sensor Integration (SSI) framework available to interested researchers.